REMARKS

Claims 1-11 are pending in the application. Claims 1-11 are rejected. All rejections are respectfully traversed.

Claim 7 has been amended to depend from independent claim 1.

The invention extracts speech recognition features from a speech signal coded as a bitstream by decoding the bitstream to recover linear predictive coding (LPC) filter parameters and a residual signal. The linear predictive coding filter parameters and the residual signal are discriminatively combined into speech recognition features, which can then be recognized.

Applicants believe that by combining the residual signal with the LPC parameters a much better speech recognition is attainable. This combining has never been done before in the prior art, and is therefore novel.

Claims 1 and 6 are rejected under 35 U.S.C. 102 as being anticipated by Hershkovits et al., (US 6003004 - Hershkovits).

This is how Hershkovits works. Energy levels of the residual signal determine word boundaries. At this point, the residual signal is not used again. The LPC parameters are extracted only from the vocoder data between the word boundaries.

In greater detail, The Hershkovits voice recognizer operates ***only*** on extracted LPC parameters. Hershkovits uses the residual signal only to determine the energy in the signal, see column 5.

Applicants have recognized that the residual signal $\epsilon(n)$ can be utilized to estimate the energy since, as is known in the art, the residual signal describes the air pressure through the vocal tract while the LPC parameters describe the structure of the vocal tract and are, thus, generally independent of speech volume. As a result, the residual signal is highly correlated to how loudly or quietly a person talks.

The energy is only used to determine word boundaries for the recognizer to operate on the LPC parameters, see column 2.

Thus, the recognizer **22** searches (step **32** of FIG. **2**) for the start and end of a word within the energy signal. The start of a word is defined as the point **37** where a significant rise in energy begins after the energy signal has been low for more than a predetermined length of time. The end of a word

Also, see Figure 8, and energy estimation section beginning at column 7, line 39.

Hershkovits never combines the residual signal with the LPC parameters. Certainly, Hershkovits does not perform a discriminant combining as known in the art. The application give two examples of how discriminant combining can be performed, see last two paragraphs at page 10, namely Fisher's linear discriminant analysis (LDA), or using a discriminatory neural network.

Applicants assert that by not using the residual signal, the performance of the recognition is degraded.

The Examiner cites column 5, line 23 to column 6, line 18 as teaching the claimed "discriminatively combining the linear predictive coding filter parameters and the residual signal into speech recognition features."

Line 23-52 deal strictly with estimating energy. There is no discriminately combining, read:

FIG. **6** illustrates, in general form, the operations of vocoder based, voice recognizer **50** on a compressed frame such as the frame **52**.

As in the prior art, the energy of the frame is determined once the frame, in step **58**, has been received. However, in the present invention, the energy is estimated (step **60**) from the vocoder data, rather than from the sampled data, and the energy estimation does not involve reconstructing the sampled data.

Applicants have recognized that the residual signal $\epsilon(n)$ can be utilized to estimate the energy since, as is known in the art, the residual signal describes the air pressure through the vocal tract while the LPC parameters describe the structure of the vocal tract and are, thus, generally independent of speech volume. As a result, the residual signal is highly correlated to how loudly or quietly a person talks.

In accordance with a preferred embodiment of the present invention, one method of estimating the energy is to determine the energy in the residual signal, per frame, or, if the frames are divided into subframes, per subframe. Mathematically, this can be written as:

$$\tilde{E}_i = \sum_{n=1}^{M} \epsilon(n)^2 \qquad \text{Equation 2}$$

where $\tilde{E}_i$ is the energy in the ith frame, the residual signal $\epsilon(n)$ is reconstructed from the vocoder data and the number M is the number of sample points in the frame or subframe.

The remainder of the column, lines 53-67 deals further with energy estimation and searching for word boundaries, as explained above, there is no combining, read:

FIG. **7** is not a replica of the energy signal of FIG. **3**. However, the estimated energy signal is highly correlated with the prior art energy signal. The start and end points for the signal of FIG. **7**, labeled **62** and **63**, respectively, are also at about 0.37 sec and 0.85 sec, respectively.

Other methods of estimating the energy from the vocoder data are incorporated in the present invention, some of which are described hereinbelow.

Returning to FIG. **6**, the vocoder based, voice recognizer **50** searches (step **64**) for word boundaries in the estimated energy signal. If desired, voice recognizer **50** can refine the location of the word boundaries by using any of the char-

Column 6, lines 1-7 deals only with extracting LPC parameters from sections within word boundaries, as identified above. There is no discriminately combining, read:

acteristics of the LPC parameters (such as their mean and/or variance) which change sharply at a word boundary.

If a word is found, as checked by step **66**, recognizer **50** extracts (step **68**) the LPC word parameters from the vocoder data. Step **68** typically involves decoding the encoded LPC parameters provided in voice compression frame **52** and converting them to the LPC coefficients.

Then, for the remainder of the cited section, lines 8-18, describes how the recognition features are strictly based on extracted LPC parameters. There is no mention of the residual signal her, nor is there any type of discriminately combining, read:

Recognizer **50** then calculates (step **70**) its recognition features from the extracted LPC coefficients. These recognition features can be any of the many LPC based parameters, such as cepstrum coefficients, MEL cepstrum coefficients, line spectral pairs (LSPs), reflection coefficients, log area ratio (LAR) coefficients, etc., all of which are easily calculated from the LPC coefficients. Thus, if the vocoder uses one type of LPC parameter and the recognizer **50** use another type of LPC parameter, recognizer **50** can convert from one to the other either directly or through the LPC coefficients.

The Examiner is respectfully requested to point out what lines in the above sited sections describe discriminately combining LPC parameters with the residual signal.

The sections never describe combining, or anything equivalent to combining, such as uniting, joining, mixing, merging, or coalescing. Upon reading it should be crystal clear that the recognition features can only be cepstrum coefficients, MEL cepstrum coefficients, line spectral pairs (LSPs), reflection coefficients, log area ratio (LAR) coefficients, which are easily **calculated from the LPC coefficients**.

The residual signal is only used for finding the word boundaries where the LPC parameters are evident.

As stated above, Hershkovits only computes the *energy* of the residual signal. Claimed is analyzing an entire *spectrum* of the residual signal. One of ordinary skill in the art would never confuse 'energy', which we all know to mean power, with spectrum, which relates to signal frequency.

Furthermore, with all due respect, the Examiner's comparison makes no sense. A frame is 5 to 20 msec worth of samples. A frame was and never will be a spectrum. Disclosing the processing of frame samples does not anticipate analyzing the spectrum of a residual signal as claimed.

Claims 2-4 are rejected under 35 U.S.C. 103(a) as being unpatentable over Hershkovits et al., in view of Aguilar et al. (US 6691082 - Aguilar).

Aguilar upsamples the raw speech signal from 4 KHz to 8 KHz, not the LPC parameters themselves as claimed. Note also that the claimed cepstral vectors are obtained from *upsampled LPC parameters*, and not from an upsampled acoustic signal as in the recited references.

Claims 2-4 are rejected under 35 U.S.C. 103(a) as being unpatentable over Hershkovits et al., in view of Park (US 6,108,624).

Park pads positions of a time axis corresponding to a second subframe with zeros, initializes the pitch filter and LPC filter to zero. Claimed are setting *short-term prediction coefficients* to zero. Short-term prediction coefficients are not positions on a time axis, nor pitch or LPC filters.

The application merely states that a 32-dimensional log spectra is derived from the residual signal. The specification at page 11 does not say "residual log-spectra **must** be derived before inputting into the neural network," this is an inaccurate characterization of the invention by the Examiner. The prior art cited at page 11 only has to do with speech recognition, generally. There is no admission or indication that the prior art teaches "deriving a high-dimensional log spectra from *up-sampled LPC parameters*," as stated in claim 7. The Examiner is requested to consider all limitations in the claim. As stated above with respect to claim 2, Applicants believe that upsampling LPC parameters is novel. The prior art upsamples acoustic signals, and then derives LPC parameters.

Claims 8-10 are rejected under 35 U.S.C. 103(a) as being unpatentable over Hershkovits et al., in view of Kuhn (US 6,343,267).

At column 9, Kuhn discloses:

> FIG. **5** shows how the maximum likelihood technique works. The input speech from the new speaker is used to construct supervector **70**. As explained above, the supervector comprises a concatenated list of speech parameters, corresponding to cepstral coefficients or the like. In the illustrated embodiment these parameters are floating point numbers representing the Gaussian means extracted from the set of Hidden Markov Models corresponding to the new speaker. Other HMM parameters may also be used. In the illustration these HMM means are shown as dots, as at **72**. When fully populated with data, supervector **70** would contain floating point numbers for each of the HMM means, corresponding to each of the sound units represented by the HMM models. For illustration purposes it is assumed here that the parameters for phoneme "ah" are present but parameters for phoneme "iy" are missing.
>
> The eigenspace **38** is represented by a set of eigenvectors

Specifically, Kuhn concatenates speech parameters such as cepstral coefficients with themselves. In contrast, the invention concatenates cepstral vectors with high-dimensional log spectra.

The claimed invention reduces the dimensionality of an extended vector that is a concatenation of a cepstral vector and a high-dimensional log-spectra derived from a residual signal. Applicants firmly believe this is novel. Applicants are unaware of any prior art that describes this combination of limitations. The reference cited at page 10 and 11 are unrelated to what is specifically claimed.

It should be noted that the application makes it clear that "the invention enables the design of a distributed speech recognition system where feature extraction need not be performed on a user's handheld device. This reduces the immediate to change existing coding and transmission standards in telephone networks. It should also be understood, the invention makes the type of codec used transparent to the speech recognizer, which is not the case when the features are extracted from a reconstructed bitstream."

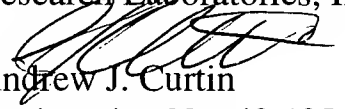
It is well known that the typical distributed system is in terms of a client/server model. In the instant application, the client is a handheld communications device, such as a cell phone, and the server is operated by the service provider, e.g., the telephone company. In this scenario, it is desired to simplify the cell phone, and have the speech recognition done at the server. Up to now to now, devices that perform speech recognition do both the feature extraction and the recognition based on the extracted feature. In contrast, the invention does the extraction at the client, the cell phone, and the recognition itself at the server. It is this unexpected division

of labor that provides advantages to applications designed according to the invention, particularly in a system with a distributed architecture.

In view of the foregoing, it is respectfully submitted that the application is in condition for allowance and an early indication of the same is courteously solicited. The Examiner is respectfully requested to contact the undersigned by telephone at the below listed telephone number, in order to expedite resolution of any remaining issues and further to expedite passage of the application to issue, if any further comments, questions or suggestions arise in connection with the application.

To the extent necessary, a petition for an extension of time under 37 C.F.R. 1.136 is hereby made. Please charge any shortage in fees due in connection with the filing of this paper, including extension of time fees, to Deposit Account 50-0749 and please credit any excess fees to such deposit account.

Respectfully submitted,
Mitsubishi Electric
Research Laboratories, Inc.



Andrew J. Curtin
Registration No. 48,485

201 Broadway, 8th Floor
Cambridge, MA 02139
Telephone: 617-621-7573
Facsimile: 617-621-7550